Research Article | Vol. 13, No. 4 / 1 Apr 2022 / Biomedical Optics Express 2018
Biomedical Optics EXPRESS

Improving the generalization of glaucoma detection on fundus images via feature alignment between augmented views

CHENGFENG ZHOU,[1] JUAN YE,[2] JUN WANG,[3] ZHIYONG ZHOU,[4] LINYAN WANG,[2] KAI JIN,[2] YAOFENG WEN,[1] CHUN ZHANG,[5] AND DAHONG QIAN[1,*]

[1]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Department of Ophthalmology, the Second Affiliated Hospital of Zhejiang University, Hangzhou, China
[3]Zhejiang University City College, Hangzhou, China
[4]School of Mechanic, Shanghai Dianji University, Shanghai, China
[5]Department of Ophthalmology, Peking University Third Hospital, Beijing, China
[*]dahong.qian@sjtu.edu.cn

Abstract: Convolutional neural networks (CNNs) are commonly used in glaucoma detection. Due to the various data distribution shift, however, a well-behaved model may be plummeting in performance when deployed in a new environment. On the other hand, the most straightforward method, data collection, is costly and even unrealistic in practice. To address these challenges, we propose a new method named data augmentation-based (DA) feature alignment (*DAFA*) to improve the out-of-distribution (OOD) generalization with a single dataset, which is based on the principle of feature alignment to learn the invariant features and eliminate the effect of data distribution shifts. *DAFA* creates two views of a sample by data augmentation and performs the feature alignment between that augmented views through latent feature recalibration and semantic representation alignment. Latent feature recalibration is normalizing the middle features to the same distribution by instance normalization (IN) layers. Semantic representation alignment is conducted by minimizing the Topk NT-Xent loss and the maximum mean discrepancy (MMD), which maximize the semantic agreement across augmented views from individual and population levels. Furthermore, a benchmark is established with seven glaucoma detection datasets and a new metric named mean of clean area under curve (*mcAUC*) for a comprehensive evaluation of the model performance. Experimental results of five-fold cross-validation demonstrate that *DAFA* can consistently and significantly improve the out-of-distribution generalization (up to +16.3% *mcAUC*) regardless of the training data, network architectures, and augmentation policies and outperform lots of state-of-the-art methods.

1. Introduction

In recent years, the convolutional neural networks have outperformed human experts across a plethora of applications [1,2], and have been a popular method for glaucoma detection [3–5]. However, models performing well under the independent and identically distributed assumptions are not enough in real clinical implementation. CNNs are highly vulnerable to data distribution shifts in new deployment environments, which are inevitable due to the various imaging equipment, the different operating procedures, and the variation in annotation protocols [6,7]. A common solution to address the data distribution shifts is collecting more data, but it is expensive, time-consuming, and even impractical, especially in clinical scenarios. Therefore, it is desirable to train the model with the data from a certain distribution and make it robust to the unforeseen data distribution shifts between deployment environments.

#450543 https://doi.org/10.1364/BOE.450543
Journal © 2022 Received 10 Dec 2021; revised 3 Feb 2022; accepted 27 Feb 2022; published 11 Mar 2022

Biomedical Optics EXPRESS

Most previous studies in the field of medical imaging analysis improved the CNNs robustness using multiple datasets collected from various sites. For example, Liu et al. [8] enhanced prostate segmentation by multi-site-guided knowledge transfer. Bateson et al. [9] trained the segmentation networks with the constraint of domain-invariant prior knowledge (e.g., the size or shape of segmentation regions). Chen et al. [10] proposed an unsupervised domain adaptation framework that aims to align the source data and the target data both in the image and the feature perspectives. The main idea of the abovementioned methods is to dig the shared information across different datasets and extracting universal representations that robust to the data distribution shifts. Despite the promising performance of these methods, the requirement for multiple datasets hinders their real-world application.

On the other hand, there are many feasible methods under the setting of single dataset in natural image analysis [11–21]. For example, Tang et al. [21] designed two normalization techniques, namely SelfNorm and CrossNorm. The former recalibrates the channel-wise statistics in an attention manner, while the latter exchanges the statistics between the middle features. Hendrycks et al. [16] developed a data augmentation technique called AugMix that incorporates multiple augmented images as one image in the MixUp manner [14]. However, the effectiveness of methods developed on natural images is unknown on medical images, considering their special characterizations such as fine-grained classes, large intra-class variations, small inter-class variations, few samples, etc.

In this study, we propose a novel method called *DAFA* to improve the out-of-distribution generalization with a single dataset for glaucoma detection in fundus images. The *DAFA* roots from the hypothesis in feature alignment (FA) methods [22–25] that robust representations should share a same feature space regardless of the data distribution shifts. Accordingly, we aim to learn a robust feature via feature alignment between two different distributions that are simulated by generating sample variants with stochastic data augmentation. Specifically, we use the instance normalization layers without parameters [26] to recalibrate the distribution of middle features. In the meantime, we maximize the identity agreement between the latent representations of the same sample by minimize the Topk NT-Xent loss and the maximum mean discrepancy between two batch representations under different augmented views. This mechanism can learn a view-agnostic representation (i.e., robust to distribution shifts) effectively. Thus, the learnt decision-boundary in feature space could be consistent across the various test data.

Our contributions are summarized as follows:

- We propose the *DAFA* method to improve the OOD generalization with a single dataset in glaucoma detection. Through a novel feature alignment strategy, robust features can be learned to handle the data distribution shifts.
- To comprehensively and reliably evaluate the model robustness, we propose a new evaluation metric, i.e., mean of clean area under curve (*mcAUC*) based on seven glaucoma detection datasets. This metric provides a simulation of real applications and a comprehensive OOD generalization evaluation without the inherent dataset bias [20].
- Extensive experimental results demonstrate that our method consistently and substantially improves the OOD generalization performance regardless of training data, the model architecture, and augmentation policies and significantly outperforms most existing OOD generalization methods.

2. Related work

A straightforward solution to improve the OOD generalization performance is to collect some data from the target domain and re-training the model in a domain adaptation [22] or transfer learning manner [27]. However, it is unfeasible when the target data are inaccessible. Fortunately,

Biomedical Optics EXPRESS

domain generalization (DG) [22–25,28] can be applied in the cases where multi-source data exist. Its primary aim of DG is minimizing the representation mismatch between domains and maximizing the separability of data. For instance, Motiian et al. [23] proposed a semantic alignment loss and a separation loss to achieve this goal. Carlucci et al. [28] captured the sharing knowledge crossing the domains by solving a jigsaw puzzle. However, as aforementioned, collecting multi-sources data is usually as impractical as collecting target domain data, especially in clinical practice due to high cost, sensitive information protection, data scarcity, etc.

To avoid the issues of data collection, several methods have been proposed to improve the OOD generalization with a single dataset. These methods can be mainly divided into two categories as follows: 1) data-based methods and 2) model-based methods.

The data-based methods address the problem through data augmentation or data pre-processing. It is a common sense that traditional data augmentations (e.g., translation, rotation, scaling, flipping, etc.) can improve the model robustness. Several advanced techniques including automatic augmentations [13], mixed samples augmentations [14–16], adversarial examples augmentation [17], and style transfer augmentation [18] also were demonstrated the same property. Besides, a proper data pre-processing can greatly eliminate the distribution shifts. For example, N4BiasFieldCorrection [29] is a conventional routine for MR images analysis. Standardizing samples with Contrast Limited Adaptive Histogram Equalization (CLAHE) can improve the classification of corrupted images [20].

On the contrary, the model-based methods focus on designing special network architectures to reduce the impacts of distribution shifts. For example, Pan et al. proposed the IBN-Net [11] which integrates instance normalization and batch normalization as the building blocks, and applied it to cross-domain generalization. Zhang [12] made the CNNs shift-invariant by combining low-pass filtering with anti-alias. Paul et al. [30] demonstrated the surprising robustness of the Vision Transformer (ViT).

In contrast, the *DAFA* method is inspired by DG and incorporates the advantage of both data-based and model-based methods. *DAFA* can fully leverage the potential of a single dataset which may capture only a narrow slice of the entire distribution of real data. It is more feasible for the scenarios where the target data is unforeseeable and the multi-source datasets are non-existent.

3. Methods

Given a set of N samples $(X, Y) = \{(x_0, y_0), (x_1, y_1), \ldots, (x_{N-1}, y_{N-1})\}$ which are drawn from training distribution $\mathcal{P}_{train}(X, Y)$, the problem is find the optimal model $\mathcal{F}_\theta^*$ which can generalize best on test distribution $\mathcal{P}_{test}(X, Y)$. It can be described as:

$$\mathcal{F}_\theta^* = \arg\min_{f_\theta} \mathbb{E}_{X,Y\sim P_{test}} \left[\mathcal{L}\left(f_\theta(X), Y\right)\right], \tag{1}$$

where $\mathcal{L}\left(f_\theta(X), Y\right)$ is the loss function for the model optimization.

In real world applications, the independent and identically distributed assumption barely holds, namely $\mathcal{P}_{train}(X, Y) \neq \mathcal{P}_{test}(X, Y)$. The source of training data and test data may be sourced from the different environments ϵ. Here, we denote the training data and the test data as $(X^\epsilon, Y^\epsilon) \sim \mathcal{P}^\epsilon$ and $(X^{\varepsilon_i}, Y^{\varepsilon_i}) \sim \mathcal{P}^{\varepsilon_i}$ respectively, where the test environment ε_i is infinite and unforeseeable. The optimal model $\mathcal{F}_\theta^*$ under this context is:

$$\mathcal{F}_\theta^* = \arg\min_{f_\theta} \sum_i \mathbb{E}_{X,Y\sim\mathcal{P}^{\varepsilon_i}} \left[\mathcal{L}\left(f_\theta(X), Y\right)\right]. \tag{2}$$

Here, we decompose this optimal model $\mathcal{F}^*$ as $\mathcal{F}^* = g \circ f$ where $f : X \rightarrow R$ is a representation learning function and $g : R \rightarrow Y$ is a classifier. Previous work [22–25] has demonstrated that a invariant representation R elicits a invariant prediction for classifier g. Thus, the goal of our

Biomedical Optics EXPRESS

method is learning a invariant representation R which not be changed with data distribution shifts. The formulation is:

$$\min_{f,g} \mathbb{E}_{(x,y)\sim\mathcal{P}^{\epsilon}} \mathcal{L}(g(f(x)), y) + \lambda \ell_{\text{reg}}, \tag{3}$$

where ℓ_{reg} denotes the regularization term which is performed as feature alignment between augmented views in our method, and λ is the tradeoff parameter.

Figure 1 shows the main structure of *DAFA*. Specifically, given a mini-batch X of n samples, the augmentation policy $\mathcal{T}$ and $\mathcal{T}^{'}$ stochastically transforms it into two different augmented views (denotes by v and $v^{'}$), generating the two sets of variants $\hat{X}$ and $\bar{X}$. Then, both sets are fed into a specific CNN backbone $f(\cdot)$ to extract the respective representations, namely $\hat{R}$ and $\bar{R}$, which are finally used to predict their labels $\hat{y}$ or $\bar{y}$ via a classifier $g(\cdot)$. The cross-entropy loss $\mathcal{L}_{CE}$ for glaucoma detection will supervise this prediction process. Meanwhile, the Multi-layer Perceptron (MLP) $h(\cdot)$ maps the representations to two semantic embeddings $\hat{E}$ and $\bar{E}$, aiming to maximize the agreement between the representations through the Topk NT-Xent loss $\mathcal{L}_{Con}$ and the Maximum Mean Discrepancy loss $\mathcal{L}_{MMD}$ [31]. More details of our method are described in the following sections.

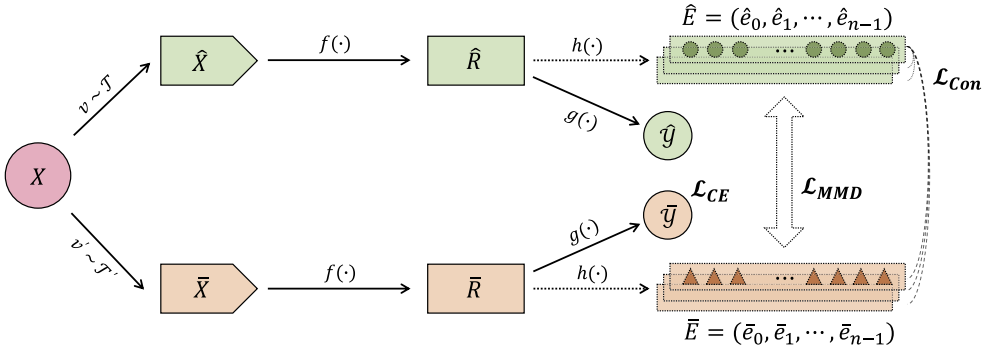

Fig. 1. Overview of *DAFA*. Given a mini-batch X of n samples, the augmentation policies $\mathcal{T}$ and $\mathcal{T}^{'}$ stochastically transforms it into view v and $v^{'}$, respectively. A CNN backbone $f(\cdot)$ extracts the representations $\hat{R}$ and $\bar{R}$ from the augmented data $\hat{X}$ and $\bar{X}$ respectively. Subsequently, a neural network $g(\cdot)$ predicts the respective label $\hat{y}$ or $\bar{y}$ for $\hat{X}$ and $\bar{X}$ according to these representations. Meanwhile, another neural network $h(\cdot)$ maps these representations to semantic embeddings $\hat{E}$ and $\bar{E}$. $\mathcal{L}_{Con}$ and $\mathcal{L}_{MMD}$ can maximize the agreement between the semantic embeddings from individual and population levels. $\mathcal{L}_{CE}$ denotes the cross-entropy loss for glaucoma detection. The dotted line indicates that this module only works during the training phase.

3.1. Augmented views

The data augmentation policy $\mathcal{T}$ and $\mathcal{T}^{'}$ should apply content preserving transformations such as cropping, resizing, mirroring, color jittering, and color dropping on the input samples X to maintain the task-relevant information [32]. Meanwhile, it should minimize the mutual information between different views as well as possible to help the model capture more generic knowledge. In our preliminary experiments, we found that the composition of multiple augmentation operations and random color distortion could considerably benefit the OOD generalization on glaucoma detection. Note that the hyperparameters for cropping operation should be carefully set to preserve most of the main pathological area (e.g., optic cup and disc, or its surrounding blood vessel and optic nerve area) for the meaningful glaucoma detection. After the augmentations, we can promote data diversity and create various augmented views stochastically.

Biomedical Optics EXPRESS

3.2. Latent features recalibration

Regardless of which backbone network $f(\cdot)$ (e.g., ResNet50 [33], ResNeXt50 [34], and DenseNet121 [35]) is applied, we replace the batch normalization (BN) [36] with instance normalization [26] to realize the latent features recalibration.

Given a feature $F \in \mathbb{R}^{N\times C\times H\times W}$, the standard IN performs the normalization as follows:

$$\hat{F} = \gamma\frac{(F-\mu)}{\sigma} - \beta, \tag{4}$$

where μ and σ represent the mean and standard deviation along the spatial dimension of each channel (i.e., $H \times W$) , and γ and β are learnable parameters for scale and shift, respectively. Here, we remove the γ and β in Eq. (4), thus the IN layers not only eliminate instance-specific style discrepancy, but also recalibrate the feature of each channel to the same Gaussian distribution $\mathcal{N}(0,1)$ and prevent the outputs from being dominated by some specific channels [37]. Besides, IN can avoid the statistics inconsistent at training and testing time, because IN uses the batch statistics instead of the running estimates [38]. In conclusion, IN ensure both high OOD generalization and discrimination capability of the CNN models.

3.3. Semantic representation alignment

The classifier $g : \mathbb{R}^c \to \mathbb{R}^2$ is a fully connected layer to make a prediction for the input sample according to the element $r \in \mathbb{R}^c$ of representations $\hat{R}$ or $\bar{R}$. Multi-layer Perceptron $h : \mathbb{R}^c \to \mathbb{R}^{128}$, also called the projection head, is designed for the semantic representation alignment task. Here, we denote the element of $\hat{E}$ or $\bar{E}$ as e. Using $h(\cdot)$ to project the representation r as the semantic embedding e can eliminate the semantic discrepancy between augmented views and avoid learning the task-irrelevant feature.

Given a set $\hat{E} \cup \bar{E} = \{e_i |\ i \in I\}$ where $I \equiv \{0,\ 1,\ \cdots,\ 2n-1\}$, we let $j(i)$ be the index of the another semantic embedding derived from the same sample x_i. e_i and $e_{j(i)}$ are the normalized projection of x_i under two different augmented views. We define the pair $(e_i, e_{j(i)})$ as positive pair and (e_i, e_k) with $k \in I\backslash\{i,\ j(i)\}$ as negative pairs. The Topk NT-Xent loss $\mathcal{L}_{Con}$ only computed across positive pairs set $A(i)$ is defined as follows:

$$\mathcal{L}_{Con} = \frac{1}{2n}\sum_{i\in I}\ell_i = -\frac{1}{2n}\sum_{i\in I} log\frac{exp(e_i \cdot e_{j(i)}/\tau)}{\sum_{a\in A(i)} exp(e_i \cdot e_a/\tau)}, \tag{5}$$

where loss ℓ_i is the loss for the positive pair $(e_i, e_{j(i)})$; the samples similarity is conduct by the inner product, namely the cosine similarity; τ denotes a temperature parameter; and the set $A(i) \equiv \{a \in Topk(i, I, \mathcal{K})|\ a \neq i\}$ where $Topk(i, I, \mathcal{K})$ returns the indexes of $\mathcal{K}$ most similar semantic embeddings of e_i. The aim of ℓ_i is to maximize the agreement between two semantic embeddings of the same sample (i.e., maximizing the numerator in Eq. (5), $e_i \cdot e_{j(i)} \to 1$) and minimizing that in negative pairs (i.e., minimizing the denominator in Eq. (5), $e_i \cdot e_a \to 0$). Thereby, more similar negative pairs or less similar positive pairs implies harder to optimize and the challenge for Topk NT-Xent loss $\mathcal{L}_{Con}$ will increase as the batch size n grows. Compared to the NT-Xent loss [32], replacing the denominator indexes set $\bar{A}(i) \equiv \{a \in I\backslash i\}$ with $A(i) \equiv \{a \in Topk(i, I, \mathcal{K})|\ a \neq i\}$ can directly increase the contribution of hard negatives and decrease that of easy negatives.

With the Topk NT-Xent loss $\mathcal{L}_{Con}$, the model will project the variants of a sample to the same semantic point according to their own identity discriminative information, which effectively eliminates the effect of transformations or perturbations (i.e., data distribution shifts) in individual level.

MMD loss is widely used in domain adaptation [39] and domain generalization [22,23], and provides a criterion for estimating the distance between distributions without the requirement of

Biomedical Optics EXPRESS

the intermediate density estimation. The definition of $\mathcal{L}_{MMD}$ is:

$$\mathcal{L}_{MMD} = |\frac{1}{n}\sum_{i=0}^{n-1}\phi(\hat{e}_i) - \frac{1}{n}\sum_{i=0}^{n-1}\phi(\bar{e}_i)|_{\mathcal{H}}^2, \tag{6}$$

where $\mathcal{H}$ represents a Reproducing Kernel Hilbert Space (RKHS) with Gaussian kernel $k(e, e^{'}) = exp(-\frac{1}{b}\parallel e - e^{'} \parallel_2^2)$ and $\phi : R^{128} \rightarrow \mathcal{H}$ maps the features to RKHS. Minimizing the $\mathcal{L}_{MMD}$ between different views is helpful to incorporate the invariance to semantic embeddings in population level.

3.4. Object function

The overall object function $\mathcal{L}_{overall}$ is defined as

$$\mathcal{L}_{overall} = \alpha\mathcal{L}_{CE} + \beta\mathcal{L}_{Con} + \gamma\mathcal{L}_{MMD}, \tag{7}$$

where cross-entropy loss $\mathcal{L}_{CE} = \frac{1}{2n}\sum_{i=0}^{2n-1} y_i \cdot \log g(r_i)$ assesses the glaucoma detection with the ground-truth y, and $\mathcal{L}_{Con}$ and $\mathcal{L}_{MMD}$ align the semantic embeddings in different views at the individual and population levels. In our experiments, the hyperparameters α, β, and γ in $\mathcal{L}_{overall}$ are both empirically set to 1.0 and the bandwidth b in Gaussian kernel $k(e, e^{'})$ of $\mathcal{L}_{MMD}$ is set to median pairwise squared distances on training data [39].

4. Experiments

4.1. Datasets

Seven datasets are included in our benchmark as follows: LAG [5], ODIR [42], ORIGA^{-light} [43], REFUGE [44], RIMONE-r2 [45], BY, and ZR. The details of each dataset are summarized in the Table 1. Two private datasets, i.e., BY and ZR, are collected from Peking University Third Hospital and the Second Affiliated Hospital of Zhejiang University, respectively, and labeled by qualified glaucoma specialists. Before the experiments, the images are resized according to their content scale and their values are normalized to zero mean and unit variance. As shown in Fig. 2(a), the data heterogeneity on appearance and contrast apparently exists across the seven datasets. Figure 2(b) visualizes the features extracted by ResNeXt-101 32x32d WSL [41] using t-SNE. It can be noted that ORIGA^{-light} and RIMONE are far apart from the other five datasets and the REFUGE aggregates in a small region. In additional, these datasets collected from related but distinct domains can also support other research such as Domain Generalization, Multi-site Learning, Incremental Learning, etc.

Table 1. Details of the datasets in our benchmark.

Dataset	Images (pos/neg)	Camera	Resolution
LAG [5]	4854 (1711/3143)	Zeiss, Canon, and Topcon	500 × 500
ODIR [42]	3430 (325/3105)	Zeiss, Canon, and Kowa	599 × 639 to 2847 × 4248
ORIGA^{-light} [43]	650 (168/482)	unknown	1995 × 2087 to 2047 × 2597
REFUGE [44]	400 (40/360)	Zeiss Visucam 500	2056 × 2124
RIMONE-r2 [45]	455 (255/200)	Nidek AFC210	290 × 290 to 1375 × 1654
BY	4388 (873/3515)	Canon CR-2 AF	1584 × 1260 to 2303 × 2142
ZR	1839 (971/868)	Topcon TRC-NW8	2014 × 1667 to 2135 × 2713

Biomedical Optics EXPRESS

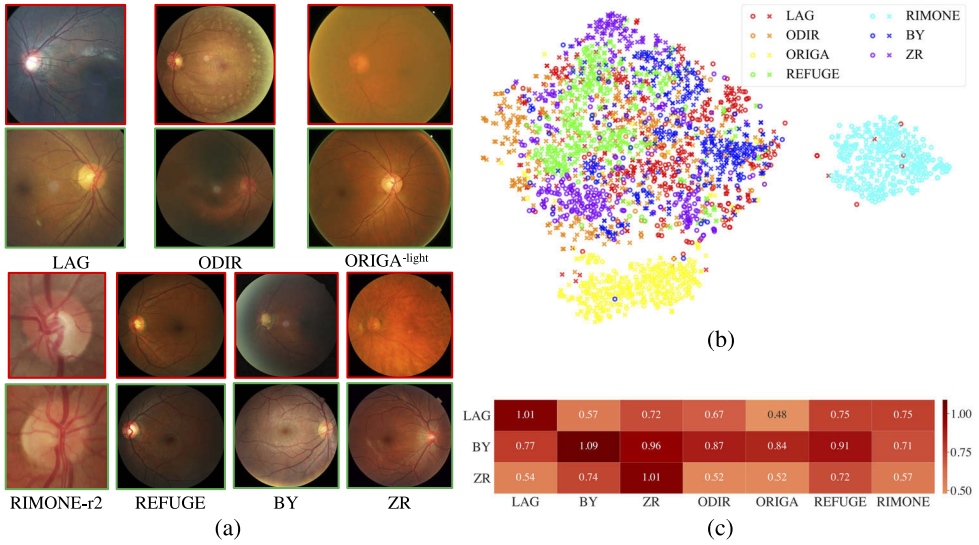

Fig. 2. (a) Visualization of samples in our benchmark. The glaucoma samples and normal samples are marked as red boxes and green boxes, respectively. (b) Visualizing the samples using t-SNE [40]. The high-dimensional features extracted by ResNeXt-101 32x32d WSL [41] are reduced their dimension by t-SNE. (c) Cross-validation results. Columns denote the source dataset, and rows denote the target datasets (report in *cAUC*).

4.2. Implementation details

In this paper, all experiments are conducted on two NVIDIA GeForce GTX 2080Ti with PyTorch implementation. Due to the data size and balanced categories, we use LAG, BY, or ZR as the source dataset, and the remaining datasets as the target datasets. The model with the maximum the area under curve (*AUC*) on the validation set (i.e., a split of LAG, BY, or ZR) is adopted to evaluate its performance on the target datasets. The default settings for training is: SGD optimizer with Nesterov momentum of 0.9; weight decay of $1e-5$; batch size of 64; epochs of 400. The learning rate increases linearly from 0 to $3e-3$ at the first 5 epochs and linearly decays to 0 following a cosine decay schedule. The automatic mixed precision training strategy is adopted for training speedup. We use different popular CNNs as the feature extractor $f(\cdot)$, a fully-connected layer as the $g(\cdot)$, and a 2-layer MLP as $h(\cdot)$ with weight matrixes $W_1 \in \mathbb{R}^{2048\times1024}$ and $W_2 \in \mathbb{R}^{1024\times128}$ and RELU activation function. The data augmentation $\mathcal{T}$ and $\mathcal{T}'$ during the training phase is same as [46]. We adopt the above settings for all experiments unless specified otherwise.

4.3. Evaluation of the proposed model

In this section, we first compare our approach with the *Baseline* (i.e., the original ResNet50 [33], ResNeXt50 [34], or DenseNet121 [35]) using various training data, networks, and augmentation policies. And then, we investigate its effectiveness in learning the features robust to distribution shifts. Finally, we conduct the comparisons with the state-of-the-art OOD generalization methods.

4.3.1. Evaluation metric

We denote the seven datasets in our benchmark as $\mathcal{D}$. In order to comprehensively assess the performance of the models, one dataset is used as the source dataset and the remaining six datasets are used as the target datasets. The source dataset $\tilde{d}$ randomly split into five subsets $\{\tilde{d}_0, \tilde{d}_1, \ldots, \tilde{d}_4\}$ to conduct the 5-fold cross-evaluation. For each iteration of the 5-fold

Biomedical Optics EXPRESS

cross-evaluation, one split $\tilde{d}_s$ is used as the validation set to select the best model m_s and remained four subsets are used for training. Finally, the target datasets $\mathcal{D}\backslash\{\tilde{d}\}$ will be used to evaluate the OOD generalization performance of model m_s.

In this benchmark, we employ *AUC* in 5-fold cross-evaluation due to its comprehensiveness and stability. We aggregate the results on the multiple target datasets as:

$$mcAUC = \frac{1}{N}\sum_{d\in\mathcal{D}\backslash\{\tilde{d}\}} cAUC_d, \tag{8}$$

where $N = |\mathcal{D}| - 1$, $d \in \mathcal{D}\backslash\{\tilde{d}\}$ denotes the target dataset, and clean area under curve (*cAUC*) is a standardized measures based on *AUC*. Since different target datasets pose different challenges, the inherent difficulties of target dataset should be considered before the aggregation and the *AUCs* need to be standardized as *cAUC*. Here, we train a ResNet50 model with all datasets $\mathcal{D}$ as the *DeepAll*. *DeepAll* serves as a strong baseline for domain generalization [24,25], and its performance implies the generalization difficulties of target datasets. Thereby, the *cAUC* of target dataset d is standardized as follows:

$$cAUC_d = \frac{\sum_{s=1}^{S} AUC_d^{m_s}}{\sum_{s=1}^{S} AUC_d^{DeepAll_s}}, \tag{9}$$

where $S = |\{\tilde{d}_0, \tilde{d}_1, \ldots, \tilde{d}_4\}|$, s indicates which split of $\tilde{d}$ or $\mathcal{D}$ used as the validation set, and $\sum_{s=1}^{S} AUC_d^{DeepAll_s}$ indicates how difficult the target dataset d. In conclusion, our benchmark successfully avoids the bias of a single target dataset and provides a fair comparison for OOD generalization.

Taking the advantage of standardized metric, we demonstrate the impact of distribution shift between datasets in the Fig. 2(c). Three main phenomena can be observed as follows: 1) The *cAUC* plummets on out-of-distribution data. 2) Its value fluctuates on these target datasets. 3) The model trained by BY shows the best performance. Thus, our benchmark indeed simulates the real deployment scenarios. And, two important conclusions can be drawn that *the result on a single dataset is unreliable and the training data distribution is a key factor to OOD generalization.*

4.3.2. Comparison with baseline settings

We evaluate the model performance across seven datasets (see the Table 2). Our method significantly outperforms the *Baseline* on all seven datasets, and approaches the *DeepAll* on REFUGE, RIMONE-r2, and BY, and even surpasses it on LAG and ODIR.

Table 2. Experimental results of ResNet50 on seven datasets (report in *AUC*). *DeepAll* is the ResNet50 trained with all seven datasets $\mathcal{D}$. Thus, it can be regarded as the upper bound for performance comparison. *Baseline* and *DAFA* are the ResNet50 trained with LAG. See Data File 1 [47] for detailed values of each fold.

	DeepAll	Baseline	DAFA
LAG	0.980±0.002	0.978±0.004	0.981±0.003
ODIR	0.764±0.023	0.638±0.017	0.830±0.016
ORIGA^{-light}	0.926±0.067	0.724±0.007	0.773±0.013
REFUGE	0.988±0.001	0.761±0.037	0.942±0.012
RIMONE-r2	0.842±0.036	0.683±0.029	0.820±0.038
BY	0.913±0.030	0.795±0.011	0.858±0.016
ZR	0.990±0.010	0.683±0.051	0.818±0.018

Table 3 shows the summarized results of ResNet50 trained by LAG, BY, or ZR. It demonstrates that *DAFA* consistently outperforms the *Baseline*. Specially, training the ResNet50 on ZR shows

Biomedical Optics EXPRESS

the largest improvement (+16.3% *mcAUC*). Besides, we find that the increase of *mcAUC* on BY is marginal. We conjecture that BY has a good sample diversity (see the Fig. 2(b) and (c)), thus the improvement of its OOD generalization is much harder than that of LAG and ZR.

Table 3. Experimental results on LAG, BY and ZR with ResNet50 (report in *mcAUC*). See Data File 2 [48] for detailed values of each fold.

Training data	*Baseline*	*DAFA*
LAG	0.792±0.008	**0.918±0.012**
BY	0.873±0.013	**0.876±0.020**
ZR	0.718±0.017	**0.835±0.010**

It should be noted that the superiority of our method is not limited by networks or augmentation policies. Table 4 shows the generality of *DAFA* on different networks. There is 0.126, 0.079, and 0.073 increase of *mcAUC* for ResNet50, ResNeXt50, and DenseNet121 respectively. Besides, the results of different augmentation policies (e.g., BarlowTwins-style augmentation [46], SimCLR-style augmentation [32] , and Fast autoaugmentation [13]) is shown in Table 5, where our method also consistently prevails the *Baseline*.

Table 4. Experimental results on LAG with popular networks (report in *mcAUC*). See Data File 2 [48] for detailed values of each fold.

Networks	*Baseline*	*DAFA*
ResNet50 [33]	0.792±0.008	**0.918±0.012**
ResNeXt50 [34]	0.827±0.013	**0.906±0.010**
DenseNet121 [35]	0.861±0.016	**0.934±0.005**

Table 5. Experimental results of ResNet50 on LAG with various augmentation policies (report in *mcAUC*). See Data File 2 [48] for detailed values of each fold.

Augmentations	*Baseline*	*DAFA*
BarlowTwins [46]	0.792±0.008	**0.918±0.012**
SimCLR [32]	0.800±0.019	**0.896±0.004**
Fast autoaugment [13]	0.752±0.011	**0.867±0.032**

In addition to the above quantitative comparisons, the Fig. 3 provides the qualitative comparison between the *Baseline* and *DAFA* in the class activation maps (CAM) [49]. It is consistently observed on all samples from target datasets that the pathological areas of glaucoma (e.g., optic cup and disc, or its surrounding blood vessel and optic nerve area.) are correctly detected by *DAFA*. In contrast, the located pathological area of *Baseline* is scattered and inaccurate.

Noteworthy, we plot the *AUC* curves of different networks in Fig. 4. It is obvious that the *DAFA* provides a more stable *AUC* curve, and speed up the model convergence.

4.3.3. Analysis of Learned Features

To verify the effectiveness of the *DAFA* more directly, we apply the symmetric Kullback-Leibler (KL) divergence [11] and the proxy $\mathcal{A}$-distance (PAD) [51] to measure the middle feature discrepancy between the source dataset $\tilde{d}_s$ and the target datasets $d \in \mathcal{D}\backslash\{\tilde{d}\}$.

We average the output of layer l of the backbone $f(\cdot)$ across the spatial dimensions and denote the features for dataset d as $\mathbb{F}_d$. We assume that each channel of $\mathbb{F}_d$ has a Gaussian distribution, i.e., $\mathbb{F}_d^i \sim N(\mu_d^i, (\sigma_d^i)^2)$. Thus, the symmetric KL divergence of the channel i between source

Biomedical Optics EXPRESS

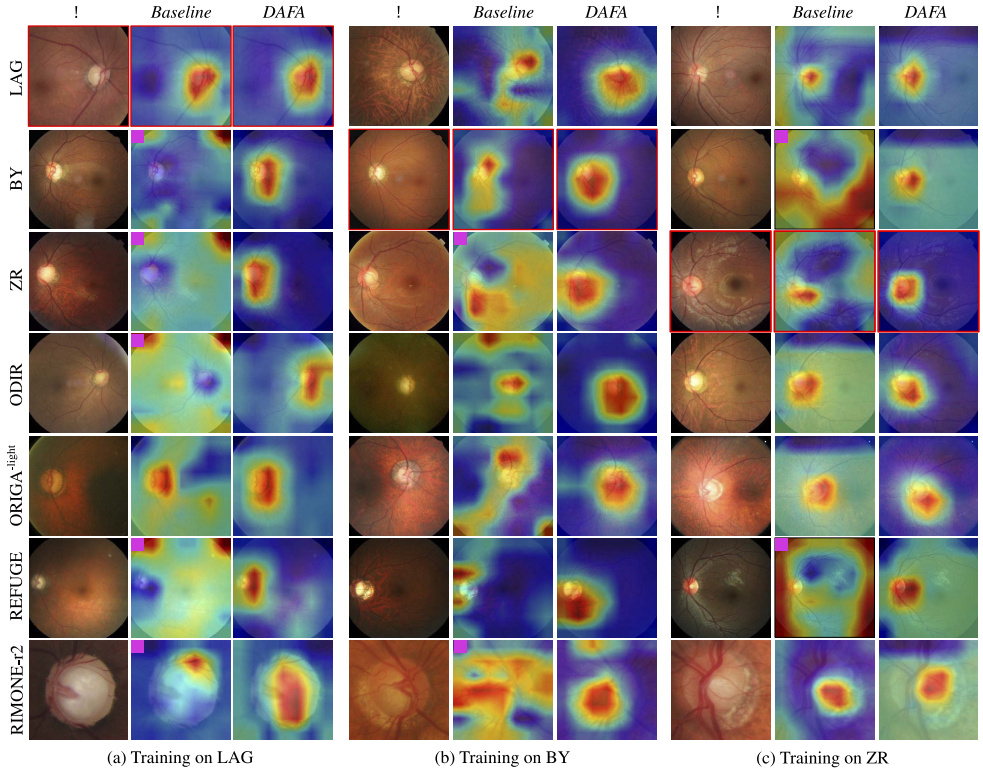

Fig. 3. Visualization results of ResNet50 models. All samples presented here are glaucoma. The red box denotes samples of the validation set, and the magenta square indicates the failed cases. The successful cases are correctly captured in the pathological areas of glaucoma (e.g., optic cup and disc, or its surrounding blood vessel and optic nerve area.). On the contrary, the heatmaps of failed cases are scattered.

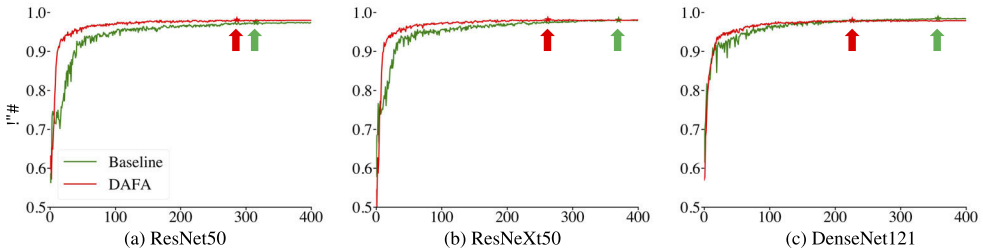

Fig. 4. The *AUC* curves of different networks trained with LAG. The horizontal axis is the epochs and the vertical axis is the *AUC* on the validation set. The highest *AUC* is marked as a star. See Data File 3 [50] for the underlying values.

Biomedical Optics EXPRESS

dataset $\tilde{d}_s$ and target dataset $d \in \mathcal{D}\backslash\{\tilde{d}\}$ is:

$$D(\mathbb{F}^i_{\tilde{d}_s} || \mathbb{F}^i_d) = KL(\mathbb{F}^i_{\tilde{d}_s} || \mathbb{F}^i_d) + KL(\mathbb{F}^i_d || \mathbb{F}^i_{\tilde{d}_s}) \tag{10}$$

$$KL(\mathbb{F}^i_{\tilde{d}_s} || \mathbb{F}^i_d) = \log \frac{\sigma^i_d}{\sigma^i_{\tilde{d}_s}} + \frac{(\sigma^i_{\tilde{d}_s})^2 + (\mu_{\tilde{d}_s} - \mu_d)^2}{2(\sigma^i_d)^2} - \frac{1}{2}. \tag{11}$$

The final feature divergence of layer l is aggregated across N target datasets and C channels as follows:

$$D_l = \frac{1}{NC} \sum_{d \in \mathcal{D}\backslash\{\tilde{d}\}} \sum_{i=1}^{C} D\left(\mathbb{F}^i_{\tilde{d}_s} || \mathbb{F}^i_d\right). \tag{12}$$

Figure 5 demonstrates the feature divergence of 17 layers in ResNet50. Apparently, *DAFA* greatly reduces the symmetric KL divergence of middle layers. An interesting find is that the *Baseline* of BY has lower feature divergence than that of LAG and ZR (especially in the deep layers), which explains why the improvement on BY is much harder than LAG and ZR.

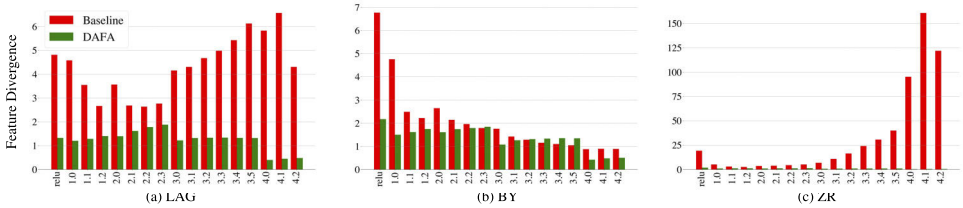

Fig. 5. Feature divergence of middle features in ResNet50. The feature divergence, namely symmetric KL divergence, is computed between the source dataset and the target datasets. See Data File 4 [52] for the underlying values.

We also reveal the discrepancy between the representations $\hat{R}$ and $\bar{R}$ in PAD. PAD is widely utilized to estimate the similarity of the source and the target representations in Domain Adaption [51]. A low PAD implies that the distribution shifts could be neglect in the feature space. As shown in the Table 6, *DAFA* significantly reduces the PAD.

Table 6. PAD of ResNet50 trained with LAG, BY, or ZR.

Training data	*Baseline*	*DAFA*
LAG	1.936	**1.842**
BY	1.942	**1.873**
ZR	1.855	**1.795**

In summary, a robust feature is learned by *DAFA* and the impacts of distribution shifts are significantly reduced in feature space. These results give us an intuition of *DAFA* that it indeed improves the OOD generalization through the feature alignment.

4.3.4. Comparison with State-of-art Methods

We implemented several state-of-the-art OOD generalization methods for comparison, including Anti-aliased ResNet50 [12] which makes model shift-invariant by integrating low-pass filtering to anti-alias, IBN-ResNet50-a and IBN-ResNet50-b [11] which address the domain or appearance variation by a well-designed IBN block, AugMix [16] which improves model robustness to unforeseen distribution shifts by mixing randomly augmented samples, AdvProp [17] which improves the image recognition model by reducing the overfitting of adversarial examples with

Biomedical Optics EXPRESS

separate auxiliary batch normalization layers, Shape-ResNet [18] which learns a shape-based representation by stylized samples, and Pertrained ResNeXt101 [53] which obtains a robust model by fine-tuning the pre-trained models.

In addition, we also report the results of several popular CNNs and domain generalization methods as a reference. Specifically, DenseNet121 [35] is a strong baseline for model robustness. EfficientNet-B0 [54] and ResNeSt50 [55] achieve the state-of-the-art performance on ImageNet dataset. The domain generalization methods including JigenDG [28], EISNet [56], and ERDG [57] both attempt to extract the domain-invariant features through the additional regularization (i.e., jigsaw puzzles, momentum metric learning, or conditional entropy maximizing).

As listed in Table 7, the models trained without specific designed data augmentations (i.e., ResNet50, DenseNet121, EfficientNet-B0, and ResNeSt50) show poor OOD generalization except for the ResNeSt50. The model robustness methods, Anti-aliased ResNet50, IBN-ResNet50-a, IBN-ResNet50-b, AugMix, AdvProp, and Shape-ResNet, can effectively improve the ResNet50, especially the IBN-ResNet50-b. And, pretraining on massive data approximately 1 billion images, even it is unrelated, could improve the model robustness significantly (i.e., Pretrained ResNeXt101). Nonetheless, our method achieves the best performance with a small dataset (i.e., 3386 images) and surpasses the above methods significantly . We conjecture that these methods are developed in the natural images without considerations for the characteristics of medical images. On the other hand, the domain generalization methods trained with the union of LAG, BY, and ZR do not outperform our method. The *mcAUC* of JigenDG, EISNet, and ERDG [57] are lower than our method by 0.041, 0.072, and 0.079, respectively.

Table 7. OOD Generalization performance of various methods trained on LAG. See Data File 5 [58] for the detailed values of each fold.

Method	*mcAUC*
DAFA	**0.918±0.012**
ResNet50 [33]	0.655±0.018
Anti-aliased ResNet50 [12]	0.712±0.024
IBN-ResNet50-a [11]	0.732±0.009
IBN-ResNet50-b [11]	0.792±0.010
AugMix [16]	0.700±0.016
AdvProp [17]	0.728±0.012
Shape-ResNet [18]	0.674±0.017
Pertrained ResNeXt101 [41]	0.918±0.021
DenseNet121 [35]	0.677±0.009
EfficientNet-B0 [54]	0.593±0.035
ResNeSt50 [55]	0.729±0.025
JigenDG [28]	0.877±0.021
EISNet [56]	0.846±0.018
ERDG [57]	0.839±0.019

Figure 6 provides the results on each target dataset. Notably, all methods do not perform well on the RIMONE-r2 and ORIGA^{-light}, because there are large content mismatches and distribution gaps between the LAG and RIMONE-r2 or ORIGA^{-light} (see Fig. 2(a) and (b)). Moverover, our method (the green curve in Fig. 6) demonstrates its superiority on each target dataset.

Biomedical Optics EXPRESS

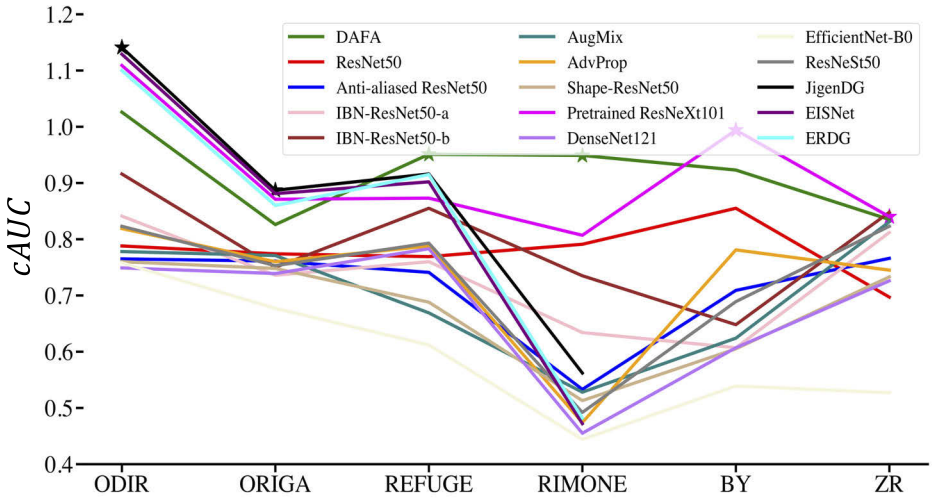

Fig. 6. The detailed results of various methods trained on LAG (report in *cAUC*). The highest *cAUCs* on every dataset are marked as a star. See Data File 5 [58] for the underlying values.

4.3.5. Ablation Studies

In this section, we analyze the influence of hyperparameters $\mathcal{K}$ and τ and which components crucially contribute to these improvements. All experiments in this section are conducted on the ResNet50 trained by LAG.

Hyperparameters influence. We analyze the influence of Topk NT-Xent loss $\mathcal{L}_{Con}$ using different values of $\mathcal{K}$ in $Topk(\cdot,\cdot,\cdot)$. As shown in Fig. 7(a), setting the $\mathcal{K}$ to 40 brings the largest gains. Then, we study which value of τ in $\mathcal{L}_{Con}$ is the best. According to Fig. 7(b), $\tau = 2.0$ is better than the other values.

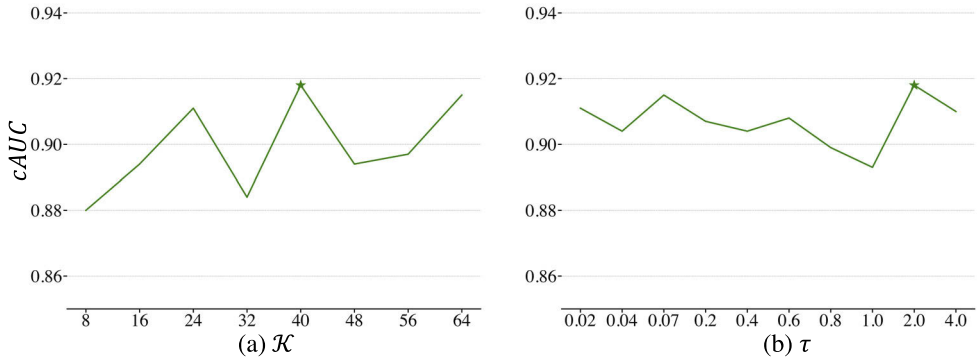

Fig. 7. Hyperparameters Tuning. The $\mathcal{K}$ (**Left**) and temperature parameter τ (**Rignt**) in Topk NT-Xent loss $\mathcal{L}_{Con}$ are adjusted according to the results of ResNet50 trained on the LAG (report in *cAUC*). See Data File 6 [59] for the underlying values.

Network design. We study how the normalization layers in $f(\cdot)$ affects the OOD generalization. We set the normalization layers to batch normalization, group normalization (GN) [38], or instance normalization with the parameters α and β (IN*param*). Then, we report their results in Table 8. Besides, we also report the result of removing the project head $h(\cdot)$. According to

Biomedical Optics EXPRESS

Table 8, we can draw the following conclusions: 1) IN surpasses the common normalization methods (i.e., BN and GN); 2) Recalibrating the feature distribution to $\mathcal{N}(0, 1)$ brings significant gains (i.e, IN vs IN*param*); 3) Project head $h(\cdot)$ is good for learning a general feature.

Table 8. Ablation results on the ResNet50 trained on LAG.

Methods	*mcAUC*
DAFA	**0.918±0.012**
w/ BN	0.885±0.018
w/ GN	0.911±0.013
w/ IN_{param}	0.903±0.009
w/o $h(\cdot)$	0.912±0.011
w/ *Supervised* $\mathcal{L}_{con}$	0.905±0.013
w/o $\mathcal{L}_{con}$	0.901±0.017
w/o $\mathcal{L}_{mmd}$	0.898±0.015

Object function design. We also discuss the object function in Table 8. We define the positive pair as two semantic embeddings of the same class in *Supervised* $\mathcal{L}_{con}$. Obviously, $\mathcal{L}_{con}$ is better than *Supervised* $\mathcal{L}_{con}$ although the latter already outperforms the *Baseline* by 0.126 *mcAUC*. Moreover, removing the object function $\mathcal{L}_{Con}$, or $\mathcal{L}_{MMD}$ decreases the *mcAUC* by 0.017, and 0.020, respectively.

5. Discussion

Fundus images have a wide variation in appearance and contrast, even the images are acquired from the same site (see Fig. 2(a)), due to the poor standardized data acquisition and individual differences. Thereby, deploying the CNN models to new sites usually brings performance drops (see Fig. 2(c)). Learning a robust glaucoma detection model is desirable in clinical applications.

In this paper, we propose a novel method called *DAFA* to improve the OOD generalization of glaucoma detection. Compared with domain adaptation or domain generalization methods, *DAFA* can fully exploit the potential of a single dataset and avoid the expensive, even infeasible data collection and the tedious training process. In addition, it significantly outperforms most existing methods that use a single dataset to improve OOD generalization and a lot of domain generalization methods that apply multiple datasets (see Table 7). Thus, *DAFA* is well-suited for the various real application scenarios. Experimental results demonstrate that our method can improve the OOD generalization regardless of the training data distribution, the model architecture, and the augmentation policies (see Table 3, Table 4, and Table 5). Another advantage of *DAFA* is that it can speed up the model convergence and stabilize the training (see Fig. 4).

In the *DAFA*, the improvement for OOD generalization stems from the Feature Alignment. Unlike the Domain Alignment [10,23], *DAFA* performs the feature alignment between two augmented views instead of two source datasets. The Feature Alignment prompts the CNN models to learn robust representations which are invariant regardless of the distribution shifts. Hence, a generic decision boundary can be learned based on these robust representations. Figure 5 and Table 6 reveal that the discrepancy between the source dataset and the target dataset is significantly reduced in the feature space, which demonstrates that the features are robust to distribution shifts. The same conclusion also can be draw from Fig. 3. The CAM visualization results demonstrate that *DAFA* correctly and consistently captures the pathological area.

Although the OOD generalization performance is improved, the limitation of our method still exists. *DAFA* slightly hampers the performance on the identically distributed data (see Fig. 4). In our future work, we plan to investigate the impact of different types of distribution shifts. For example, the distribution shifts may be caused by equipment, annotations, or population.

Biomedical Optics EXPRESS

Focusing on a certain shift could give us a clear understanding of OOD generalization. Moreover, feature normalization shows a promising improvement and brings negligible overheads (e.g., replacing the BN with IN or GN, IBN-ResNet50-a [11], and SelfNorm and CrossNorm [21]). It is worth digging into this technique thoroughly.

6. Conclusions

In this paper, we aim to learn a robust model using a single dataset. We propose a novel method called *DAFA* to enhance the OOD generalization in glaucoma detection on fundus images. This method is derived from the feature alignment. Traditionally, feature alignment is performed between two datasets, but we extend it to a single datasets with two augmented view. To evaluate our method *DAFA*, we establish a reliable benchmark with seven datasets. Experimental results on our benchmark demonstrate that *DAFA* significantly outperforms most state-of-the-art OOD generalization methods.

Funding. Science and Technology Commission of Shanghai Municipality (20DZ2220400); National Natural Science Foundation of China (81974276).

Acknowledgments. The authors thank the Institute of Medical Robotics Shanghai Jiao Tong University for the support of this research.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are available in Data File 1 [47], Data File 2 [48], Data File 3 [50], Data File 4 [52], Data File 5 [58], and Data File 6 [59].

References

1. D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell **172**(5), 1122–1131.e9 (2018).
2. R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. Mcconnell, G. S. Corrado, L. Peng, and D. R. Webster, "Predicting cardiovascular risk factors from retinal fundus photographs using deep learning," Nat. Biomed. Eng. **2**(3), 158–164 (2018).
3. X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, "Glaucoma detection based on deep convolutional neural network," in *Proc. IEEE EMBC* (2015), pp. 715–718.
4. Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs," Ophthalmology **125**(8), 1199–1206 (2018).
5. L. Li, M. Xu, H. Liu, Y. Li, X. Wang, L. Jiang, Z. Wang, X. Fan, and N. Wang, "A large-scale database and a CNN model for attention-based glaucoma detection," IEEE Trans. Med. Imaging **39**(2), 413–424 (2020).
6. I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *ICLR* (2021).
7. J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning* (The MIT Press, 2009).
8. Q. Liu, Q. Dou, L. Yu, and P. Heng, "Ms-net: Multi-site network for improving prostate segmentation with heterogeneous MRI data," IEEE Trans. Med. Imaging **39**(9), 2713–2724 (2020).
9. M. Bateson, J. Dolz, H. Kervadec, H. Lombaert, and I. Ben Ayed, "Constrained domain adaptation for image segmentation," IEEE Transactions on Medical Imaging p. 1 (2021).
10. C. Chen, Q. Dou, H. Chen, J. Qin, and P. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," IEEE Trans. Med. Imaging **39**(7), 2494–2505 (2020).
11. X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proc. IEEE ECCV*, vol. 11208 (2018), pp. 484–500.
12. R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. ICML*, vol. 97 (2019), pp. 7324–7334.
13. S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast autoaugment," in *NeurIPS*, (2019), pp. 6662–6672.
14. H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR* (2018).
15. S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE ICCV* (2019), pp. 6022–6031.
16. D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *ICLR*, (2020).
17. C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le, "Adversarial examples improve image recognition," in *Proc. IEEE CVPR* (2020), pp. 816–825.

Biomedical Optics EXPRESS

18. R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," in *ICLR*, (2019).
19. H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy," in *NeurIPS*, (2019), pp. 8250–8260.
20. D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR* (2019).
21. Z. Tang, Y. Gao, Y. Zhu, Z. Zhang, M. Li, and D. N. Metaxas, "Crossnorm and selfnorm for generalization under distribution shifts," in *IEEE ICCV* (2021), pp. 52–61.
22. H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE CVPR*, (2018), pp. 5400–5409.
23. S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proc. IEEE ICCV* (2017), pp. 5716–5726.
24. D. Li, J. Zhang, Y. Yang, C. Liu, Y. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proc. IEEE ICCV* (2019), pp. 1446–1455.
25. Q. Liu, Q. Dou, and P. Heng, "Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains," in *Proc. MICCAI*, vol. 12262 (2020), pp. 475–485.
26. D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," CoRR **abs/1607.08022** (2016).
27. E. Gibson, Y. Hu, N. Ghavami, H. U. Ahmed, C. M. Moore, M. Emberton, H. J. Huisman, and D. C. Barratt, "Inter-site variability in prostate segmentation accuracy using deep learning," in *Proc. MICCAI*, vol. 11073 (2018), pp. 506–514.
28. F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. IEEE CVPR* (2019), pp. 2229–2238.
29. N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: improved N3 bias correction," IEEE Trans. Med. Imaging **29**(6), 1310–1320 (2010).
30. S. Paul and P. Chen, "Vision transformers are robust learners," CoRR **abs/2105.07581** (2021).
31. K. M. Borgwardt, A. Gretton, M. J. Rasch, H. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," in *Proc. ISMB* (2006), pp. 49–57.
32. T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, vol. 119 (2020), pp. 1597–1607.
33. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, (2016), pp. 770–778.
34. S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE CVPR* (2017), pp. 5987–5995.
35. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR* (2017), pp. 2261–2269.
36. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, vol. 37 (2015), pp. 448–456.
37. J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," IEEE Trans. Pattern Anal. Mach. Intell. **42**(8), 2011–2023 (2020).
38. Y. Wu and K. He, "Group normalization," Int. J. Comput. Vis. **128**(3), 742–755 (2020).
39. M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *NeurIPS* (2016), pp. 136–144.
40. L. van der Maaten and G. Hinton, "Visualizing data using t-sne," Journal of Machine Learning Research **9**(86), 2579–2605 (2008).
41. D. Mahajan, R. B. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proc. IEEE ECCV*, vol. 11206 (2018), pp. 185–201.
42. S. M. T. C. Ltd. and AIIT-PKU, "Ocular disease intelligent recognition (ODIR)," Shanggong Medical Technology Co., Ltd, 2019, https://odir2019.grand-challenge.org/dataset/.
43. Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "Origa-light: An online retinal fundus image database for glaucoma analysis and research," 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology **2010**, 3065–3068 (2010).
44. J. I. Orlando, H. Fu, J. B. Breda, K. Van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P. A. Heng, J. Kim, and J. H. Lee, "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," Med. Image Anal. **59**, 101570 (2020).
45. F. Fumero, S. Alayón, J. L. Sánchez, J. Sigut, and M. González-Hernández, "RIM-ONE: An open retinal image database for optic nerve evaluation," International Symposium on Computer-Based Medical Systems (CBMS) **2011**, 1–6 (2011)..
46. J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. ICML*, vol. 139 (2021), pp. 12310–12320.
47. C. Zhou, Y. Jun, J. Wang, Z. Zhou, L. Wang, K. Jin, Y. Wen, C. Zhang, and D. Qian, "Data File 1," figshare (2022). Retrieved https://doi.org/10.6084/m9.figshare.19294316.

Biomedical Optics EXPRESS

48. C. Zhou, Y. Jun, J. Wang, Z. Zhou, L. Wang, K. Jin, Y. Wen, C. Zhang, and D. Qian, "Data File 2," figshare (2022). Retrieved https://doi.org/10.6084/m9.figshare.19294313.
49. B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE CVPR*, (2016), pp. 2921–2929.
50. C. Zhou, Y. Jun, J. Wang, Z. Zhou, L. Wang, K. Jin, Y. Wen, C. Zhang, and D. Qian, "Data File 3," figshare (2022). Retrieved https://doi.org/10.6084/m9.figshare.19294358.
51. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," J. Mach. Learn. Res. **17**, 1–35 (2016).
52. C. Zhou, Y. Jun, J. Wang, Z. Zhou, L. Wang, K. Jin, Y. Wen, C. Zhang, and D. Qian, "Data File 4," figshare (2022). Retrieved https://doi.org/10.6084/m9.figshare.19297751.
53. D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *IEEE ICCV* (2021), pp. 8340–8349.
54. M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML* (2019), pp. 6105–6114.
55. H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. J. Smola, "Resnest: Split-attention networks," CoRR **abs/2004.08955** (2020).
56. S. Wang, L. Yu, C. Li, C. Fu, and P. Heng, "Learning from extrinsic and intrinsic supervisions for domain generalization," in *Proc. IEEE ECCV*, vol. 12354 (2020), pp. 159–176.
57. S. Zhao, M. Gong, T. Liu, H. Fu, and D. Tao, "Domain generalization via entropy regularization," in *NeurIPS*, (2020).
58. C. Zhou, Y. Jun, J. Wang, Z. Zhou, L. Wang, K. Jin, Y. Wen, C. Zhang, and D. Qian, "Data File 5," figshare (2022). Retrieved https://doi.org/10.6084/m9.figshare.19297754.
59. C. Zhou, Y. Jun, J. Wang, Z. Zhou, L. Wang, K. Jin, Y. Wen, C. Zhang, and D. Qian, "Data File 6," figshare (2022). Retrieved https://doi.org/10.6084/m9.figshare.19297757.